Bi-particle adverbs, PoS-tagging and the recognition of german separable prefix verbs

Volk, Martin ; Clematide, Simon ; Graën, Johannes ; Ströbel, Phillip

Abstract: In this paper we propose an algorithm for computing the full lemma of German verbs that occur in sentences with a separated prefix. The algorithm is meant for large-scale corpus annotation. It relies on Part-of-Speech tags and works with 97% precision when the tags are correct. Unfortunately there are multi-word adverbs with particles that are homographs with separated verb particles and prepositions. Since the usage as separated particle and preposition is much more frequent, these multi-word adverbs are often incorrectly tagged. We show that special treatment of these bi-particle adverbs improves the re-attachment of separated verb particles.

Originally published at:
Volk, Martin; Clematide, Simon; Graën, Johannes; Ströbel, Phillip (2016). Bi-particle adverbs, PoS-tagging and the recognition of german separable prefix verbs. In: KONVENS 2016, Bochum, 19 September 2016 - 21 September 2016.

Bi-particle Adverbs, PoS-Tagging and the Recognition of German Separable Prefix Verbs

Martin Volk, Simon Clematide, Johannes Graën, Phillip Ströbel
University of Zurich
Institute of Computational Linguistics
volk@cl.uzh.ch

Abstract

In this paper we propose an algorithm for computing the full lemma of German verbs that occur in sentences with a separated prefix. The algorithm is meant for large-scale corpus annotation. It relies on Part-of-Speech tags and works with 97% precision when the tags are correct. Unfortunately there are multi-word adverbs with particles that are homographs with separated verb particles and prepositions. Since the usage as separated particle and preposition is much more frequent, these multi-word adverbs are often incorrectly tagged. We show that special treatment of these bi-particle adverbs improves the re-attachment of separated verb particles.

1 Introduction

Particle verbs in German often occur with verb stem and particle split over long distances. This happens in matrix clauses when the verb is finite and occurs in present or past tense, or when the verb is in imperative form. Examples:

(1) So **wies** eine bekannte Studie der Harvard University aus dem Jahr 2007 **nach**, dass ... (EN: A well-known study by Harvard University from 2007 **proved** that ...)

(2) **Nimm** das und das **mit**. (EN: **Take** this and that **along**.)

In all other tenses and forms the particle is prefixed to the verb (e.g. ... *wie eine Studie nachwies*). Therefore the particle is often called a separable prefix. When analyzing German sentences we have to re-attach the separated prefix to the verb in order to compute the correct verb lemma. Unfortunately, Part-of-Speech taggers (like the TreeTagger) assign the lemma locally and do not consider the long-distance dependency between the verb and the prefix. Hence, we need to correct the verb lemma after PoS tagging. In example 1, the PoS tagger will assign the lemma *weisen* (EN: to point) to the past tense verb form *wies*. Only the re-attachment of the prefix will lead to the correct lemma *nach+weisen* (EN: to prove) and thus to the correct meaning of the verb.

Some annotated corpora of German leave the re-attachment of separated verb prefixes open. For example, the German TIGER treebank marks only the lemma of the finite verb as in figure 1. Since the finite verb and the separated prefix are children of the same mother node S, the prefix can be assigned unambiguously to the verb. Still, this makes querying the treebank for verbs with separable prefixes a complex undertaking. However, recent versions of the TüBa-D/Z treebank do contain verb lemmas with re-attached prefixes (Versley et al., 2010). These lemmas are represented in the same way as the lemmas of the corresponding verbs in unseparated form (e.g. *nach#weisen*).

We work on the annotation of large corpora for linguistic research and information extraction. Therefore we have developed an efficient and robust algorithm to compute the lemmas of German verbs that occur with separated prefixes. In this paper we will present the algorithm. We will then argue that multi-word adverbs cause some confusion to the PoS tagger and thus require special treatment. The correct handling of these adverbs, in return, improves the precision of the re-attached lemmas.

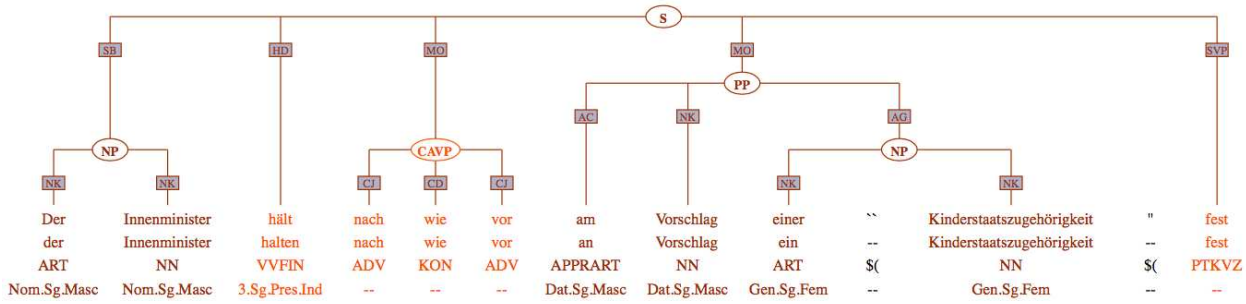

Figure 1: German syntax tree with separated verb prefix (*hält ... fest*) and multi-word adverb (*nach wie vor*) from the TIGER treebank. The multi-word adverb is annotated as coordinated adverbial phrase (CAVP). (English translation: The Interior Minister still maintains the proposal of a children citizenship.)

2 The Re-attachment Algorithm

We re-attach the separated prefix to the verb with the following algorithm. After Part-of-Speech tagging we search for a separated verb prefix (tagged as PTKVZ) and the most recent preceding finite full verb (VVFIN) or imperative verb (VVIMP) in the same sentence. In order to increase the precision we also check whether the re-combined prefix + verb lemma occurs in our corpora and is licensed by the morphology analyzer GerTwol. In this way we have compiled a list of 8500 separable German verbs.

German auxiliary verbs and modal verbs do not take separable prefixes. This means that the auxiliary verbs *haben* (EN: have) and *werden* (EN: become) must be interpreted as full verbs when they take a separable prefix. Consider for instance the verb *innehaben* in *er hat ein Amt inne* (EN: he holds an office). Other examples are *vorhaben* (EN: to intend), *fertigwerden* (EN: to be done with), or *loswerden* (EN: to get rid off).

Similarly, the modal verb *müssen* functions as full verb in combination with the prefix *durch* resulting in *durchmüssen* (EN: to have to go through), and *können* functions as full verb in *wegkönnen* (EN: to be able to leave). Our re-attachment algorithm needs to account for these cases even though state-of-the-art PoS taggers for German label all occurrences of *haben* and *werden* as auxiliary and all occurrences of *müssen* and *können* as modal verbs. Therefore we include PoS correction in the re-attachment of separated verb prefixes for these cases.

This is different from the treatment of these auxiliary and modal verbs in the TüBa-D/Z treebank. The treebank includes the re-attachment of the separated prefixes but leaves the PoS tags unchanged. This means, that in the TüBa-D/Z treebank the verb *innehaben* is a finite full verb, when the prefix is attached, but it is an auxiliary verb, when the prefix is separated. We consider this a misleading inconsistency.

Our re-attachment algorithm leads to high precision re-combined verb lemmas. We first evaluated our method against our corpus of 1.7 million German tokens from banking news (Volk et al., 2016). PoS tagging leads to a total of 9200 tokens marked as separated verb prefixes. Our algorithm re-combines 7630 prefix + verb stems (resulting in 976 types). The re-combined verbs with the highest frequencies are: *ausgehen* (345 occurrences, EN: to go out, to die down), *darstellen* (226, EN: to depict, to represent), *aussehen* (169, EN: to look like, to appear), *stattfinden* (149, EN: to take place), and *beitragen* (136, EN: to contribute). These counts do not include the occurrences of these verbs where the prefix is part of the verb form (i.e. non-separated forms): *ausgehen* (148 occurrences), *darstellen* (216), *aussehen* (106), *stattfinden* (151), and *beitragen* (292).

As a side effect we disambiguate between multiple lemma options. For example, the 3rd person singular verb form *fällt* can have the lemmas *fallen* (EN: to fall) or *fällen* (EN: to fell). The TreeTagger assigns both lemmas to this verb form. If *fällt* occurs with the separated prefix *auf*, then our re-attachment algorithm finds that only the combi-

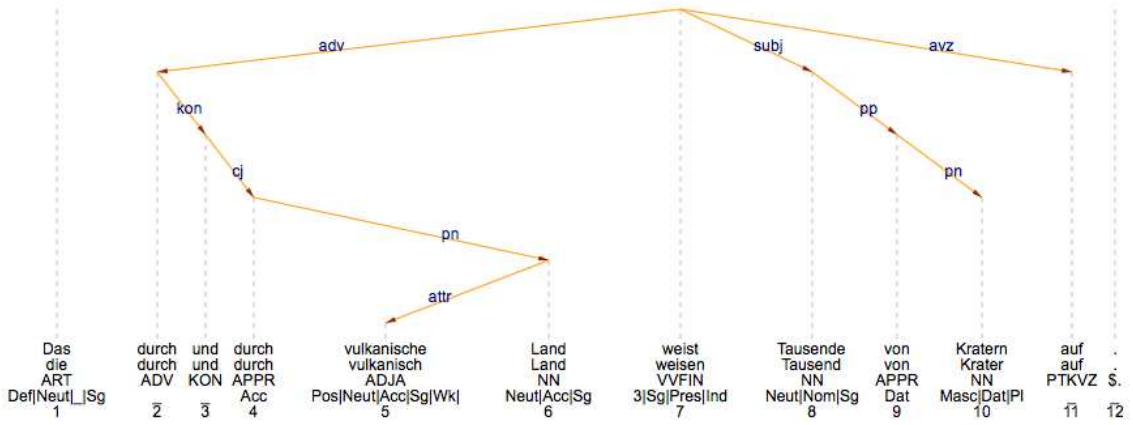

Figure 2: ParZu parser error due to incorrect recognition of the multi-word adverb *durch und durch*. (EN: The totally volcanic land has thousands of craters.)

nation *auffallen* is possible (EN: to stand out, to strike), and we eliminate the other lemma option. Obviously, this disambiguation method is dependent on the separated prefix being only acceptable with one lemma option.

We are aware of three limitations of our algorithm, all of which concern rare cases. First, the re-attachment algorithm will fail for topicalized verb prefixes that precede the finite verb. The TIGER treebank contains 33 examples of separated verb prefixes that precede the finite verb (in roughly 0.9 million tokens of manually annotated newspaper text). The topicalized prefixes in these examples are semantically heavy prefixes (e.g. *Zurück bleibt auch die Erinnerung ...*) and pronominal adverbs (e.g. *Hinzu kommt die Konkurrenz ..., Zugrunde legten die Wiesbadener ...*). We label them as adverbs and pronominal adverbs.

More serious, our re-attachment algorithm will also fail for rare cases of nested finite clauses that occur between the verb and its separated prefix. For example:

(3) Das Konsumwachstum **büsst** im Vergleich zu den vergangenen beiden Jahren, in denen die Wachstumsrate deutlich über 2,0 Prozent **lag**, markant an Schwung **ein**.
(EN: The growth in consumption considerably **loses** momentum in comparison to the past two years, in which the growth rate was clearly above 2 percent.)

This example sentence has a relative clause between the verb *büsst* and its separated prefix *ein*. Since our algorithm assigns the prefix to the most recent finite verb, it will erroneously assign it to the verb *lag* which is the finite verb in the intermediate relative clause. This problem can only be avoided by (at least) a shallow parser which detects the clause boundaries.

Thirdly, our algorithm has no provision for coordinated prefixes.

(4) In einer Deflation **nimmt** der Wert des Geldes **zu** statt **ab**, ... (EN: During a deflation the value of the money increases instead of decreases, ...)

In such examples the verb has basically two different lemmas. We could represent this in the same way as ambiguous lemmas by assigning both lemmas *zunehmen / abnehmen*, but currently this is not part of our implementation.

Other than that, if the PoS tagger recognized all verb forms and all separated prefixes correctly, then our re-attachment algorithm should work perfectly.

We evaluated our algorithm against the TüBa-D/Z treebank. Version 10 of the treebank contains a total of 9181 verb forms that have lemmas with re-attached prefixes. In the standard configuration our program correctly re-attaches 8341 separated prefixes (91%). Most of the remaining cases are verbs missing in our list of possible separable prefix verbs (which we compiled on the basis of our

alpine corpus and our banking corpus). For example *abbürsten, abwiegeln, einigeln* (EN: brush off, play down, curl up into a ball) occur in the TüBa newspaper texts but not in our list. For all these verbs, we let our morphological analyzer decide whether they are German separable prefix verbs. This adds 468 verbs to our list of acceptable separable prefix verbs and boosts the precision of our program to 96.8% re-attachments.

Some of the remaining cases are coordinations with two separated prefixes for the same verb stem (22 occurrences). Another 92 cases (roughly 1%) in the TüBa treebank are separated prefixes that precede the verb. Some are clear cases of separated prefixes (*Denn fest steht: ...; Ziemlich die Post ab geht dagegen bei ...*), many others are debatable on wether they are verb prefixes or adverbs (*Hinzu kommt, daß ...*). Only 145 prefixes (1.5%) are incorrectly attached due to a nested clause.

These numbers are computed based on manually corrected, i.e. perfect PoS tags in the TüBa-D/Z treebank. But in corpus annotation we have to rely on automatically computed PoS tags. Unfortunately, the TreeTagger has problems with the recognition of separated verb prefixes since many of them can also function as prepositions, adverbs and some other word classes. In particular, we noticed errors with the prefix *nach* (EN: after). We manually evaluated all 118 verbs with a re-attached prefix *nach* in our banking news corpus. 41 of these re-attachments (35%) were wrong.

3 Multi-word Adverbs

Closer inspection revealed that in many cases the TreeTagger had erroneously tagged an adverb or a preposition as separated prefix. We found that multi-word adverbs that are created with the coordination pattern "particle *und/wie* particle" (as e.g. *ab und zu, auf und ab, durch und durch, nach und nach, nach wie vor*; see table 1 for glosses and translations) often lead to particles that are mistakenly tagged as separated prefixes (or preposition).[1] We call this special class of multi-word adverbs **bi-particle adverbs** in analogy to the binominals as described by Gereon Müller (1997).

[1] Similar multi-word adverbs in English are *by and large, over and over, to and fro, little by little, side by side*. See also (Müller, 1997) page 3.

Mistagging of these bi-particle adverbs not only disturbs the recognition of verb lemmas but may also lead to erroneous syntax structures as in the parser output in figure 2. There, the second particle in the adverb *durch und durch* is mistagged as preposition which triggers an incorrect dependency of the following noun phrase.

For example, the PoS tagger often assigns the following tags to *nach/PTKVZ wie/KOKOM vor/APPR*, but correctly the tags should be *nach/ADV wie/KOKOM vor/ADV*. Because of these tagging mistakes we observe the following problems in the re-attachment of the separated verb prefix.

(5) Es **gibt** *nach wie vor* im deutschen Erbschafts- und Schenkungsrecht eine Privilegierung für gewerbliche Vermögen.
(EN: There is still a privilege for commercial properties in the German inheritance and donation law.)

In example 5 the TreeTagger marked *nach* as separated prefix which erroneously led to the verb lemma *nachgeben* (EN: to give in) instead of *geben* (EN: to give, there is) which does not have a separated prefix in this sentence.

(6) Schliesslich **stellen** die meisten Luxusgüterfirmen *nach wie vor* den Grossteil ihrer Produkte in Europa **her**, ...
(EN: After all, most luxury merchandise companies still produce the majority of their goods in Europe, ...)

In example 6 the same tagger error leads to the verb lemma *nachstellen* (EN: to imitate) and blocks the re-combination with the true prefix *her* into *herstellen* (EN: to produce).

(7) Wir trauen europäischen Peripherieanleihen **nach wie vor** eine gute Wertentwicklung zu.
(EN: We trust that European peripheral bonds will still have a good value development.)

In example 7 the sanity check correctly blocked the verb lemma **nachtrauen* (which does not exist), but the incorrectly tagged *nach* also blocked the re-combination of the true prefix *zu* to result in *zutrauen* (EN: to dare).

	EN glosses	EN translation	treebank freq	banking news freq	T+B corpus freq
ab und an	from and on	sometimes	3	1	10
ab und zu	from and to	sometimes	1	13	601
auf und ab	up and down	up and down	2	1	310
auf und davon	up and thereof	away	1	-	14
durch und durch	through and through	thoroughly	3	3	89
hin und wieder	to and again	sometimes	1	11	375
nach und nach	after and after	gradually	4	34	702
nach wie vor	after like before	still	62	356	396

Table 1: Multi-word adverbs with particles that also function as prepositions and separable verb prefixes. Frequencies are from the TIGER treebank (890,000 tokens, newspaper texts), from our banking news corpus (1.7 million tokens), and from our Text+Berg corpus (22.5 million German tokens).

In order to identify multi-word adverbs that contain particles which interfere with separated verb prefixes, we searched the German TIGER treebank (890,000 tokens) for coordinated adverb phrases (CADVP). There we found the bi-particle adverbs with verb prefix homographs listed in table 1. The glosses and translations prove that most of them are true multi-words whose meanings are not compositional. They contain particles that can also function as prepositions and separated verb prefixes (*ab, an, auf, durch, hin, nach, vor, zu*). Table 2 gives an overview of their tag frequencies in the treebank.

Note that table 1 is not an exhaustive list but only contains the most frequent bi-particle adverbs in our corpora. Other candidates are *aus und vorbei* (EN: clearly over), *samt und sonders* (EN: completely), *über und über* (EN: over and over).

The most frequent separated prefixes in the TIGER treebank are: *an* (669 times), *aus* (521), *ab* (433), *auf* (405), *vor* (399), *ein* (392), *zu* (244), *zurück* (227) and *mit* (220). The words *ein* and *zurück* cannot function as prepositions. Therefore we disregard them here. *mit* and *zu* are special cases since they can function as adverbs in non-conjunct constructions. *mit* can stand as adverb by itself in the sense of 'jointly' (example: *der die neue CD mit produziert hat*, EN: who has jointly produced the new CD), and *zu* functions as adverb mostly in combination with *bis* (in 121 out of the 127 cases; for example: *bis zu sechs Wochen*, EN: up to six weeks).

Since the frequencies for usages as preposition and separated prefix are much higher than the adverb usage for the particles in question, the PoS tagger is likely to mistake an adverb usage as either a preposition or verb prefix. Therefore we automatically correct the PoS tags of the multi-word adverbs (listed in table 1) in our banking corpus.

In principle, the multi-word adverbs listed in table 1 could also be coordinated prepositions or coordinated separated prefixes, except for the reduplications *durch und durch, nach und nach*. But coordinated separated prefixes are very rare and occur in word plays. Coordinated prepositions are also rare, but they still occur 24 times in the TIGER treebank. Typical examples are *mit und ohne* (EN: with and without), *in und durch* (EN: in and through), and *für und wider* (EN: for and against). It speaks for the idiomaticity of our multi-word adverbs that we have not found a single instance where they are used as coordinated prepositions.

3.1 Bi-particle Adverbs in Text+Berg

We checked how prominent the PoS tagger errors are for the bi-particle adverb *nach wie vor*. Out of 396 occurrences of this candidate in our corpus of alpine texts (the Text+Berg corpus with 22 million tokens in German), we find that *nach* is mistagged as separated prefix in 218 cases (55%), as preposition in 56 cases (14%), and even as postposition 24 times (6%). Only in 25% (98 cases) it is correctly tagged as adverb. Interestingly, in none of these 98 cases, the remainder of the multi-word adverb is correctly tagged. Some tag in this bi-particle

	preposition APPR	sep. prefix PTKVZ	adverb ADV	miscellaneous
ab	77	**433**	9	
an	**2900**	699	6	111 APZR, 1 APPO
auf	**5578**	405	3	2 APZR
aus	**2322**	521	4	65 APZR, 1 APPO
durch	**1277**	37	9	1 APPO
hin	**-**	79	63	7 APZR
mit	**6039**	220	21	
nach	**2612**	54	71	32 APPO, 1 APZR
vor	**1814**	399	67	
zu	2084	244	127	**4413 PTKZU**, 277 PTKA

Table 2: Part of Speech tag frequencies in the TIGER treebank for particles that occur in multi-word adverbs (lower case usage only). Miscellaneous PoS tags include postposition (APPO), right element of circumposition (APZR), infinitive marker (PTKZU), and adjective modifier (PTKA).

adverb is always wrong. This is clear evidence that only a special treatment or a completely different PoS tagging approach for multi-word adverbs will lead to high quality PoS tags.

Occasionally the bi-particle candidates are not multi-word adverbs. In example 8, the candidate is really a sequence of the adverb *ab* and the preposition *zu*. This is very rare. In 200 occurrences of *ab und zu* in our Text+Berg corpus we found one such occurrence.

(8) ... führen von der ursprünglich appenzellischen Weise **ab** und **zu** den Rhythmen eines ganz fremden Volkes. (EN: ... lead away from the traditional Appenzell customs and to the rhythms of a totally foreign people.)

This problem is more prominent with the candidate *ab und an* (see example 9). It occurs only 10 times in our Text+Berg corpus, but 5 of these are non-adverb cases (all predating 1925).

(9) Die Haare standen von den Köpfen **ab** und **an** der Stirne, wo das seidene Band um die Hüte ... (EN: The hair stood off from the heads and on the forehead where the silk braid around the hats ...)

Text+Berg which is a corpus with texts from the last 150 years also leads to multi-word adverbs which were prominent in the past but are no longer used, as for example **je und je** (attested 23 times from 1868 to 1958) in the meaning *always*.

(10) Von nah und fern, von dies- und jenseits des Alpengebirges sind **je und je** Geologen und Mineralogen ins Tessin gewandert, ... (EN: From near and far, from both sides of the Alpes geologists and mineralogists have always migrated to the Tissino, ...)

This multi-word adverb has been superseded by *eh und je* (EN: always) which is attested in our Text+Berg corpus 40 times since 1940.

It is striking that *nach wie vor* is the most frequent bi-particle adverb both in the TIGER treebank and in our Credit Suisse news corpus whereas *nach und nach* is the clear top frequency adverb in our Text+Berg corpus. A closer inspection revealed that this is due to the fact that the Text+Berg corpus is a collection that spans 150 years whereas the TIGER treebank and the Credit Suisse news corpus has only texts from the last 20 years. Google n-gram viewer shows that *nach wie vor* is on the upswing in recent decades whereas *nach und nach* has lost popularity during the same period (cf. figure 3).

3.2 Bi-particle Adverbs Overview

The above section on bi-particle adverbs exemplifies that many adverbs of this kind are true multi-word expressions (with non-compositional seman-

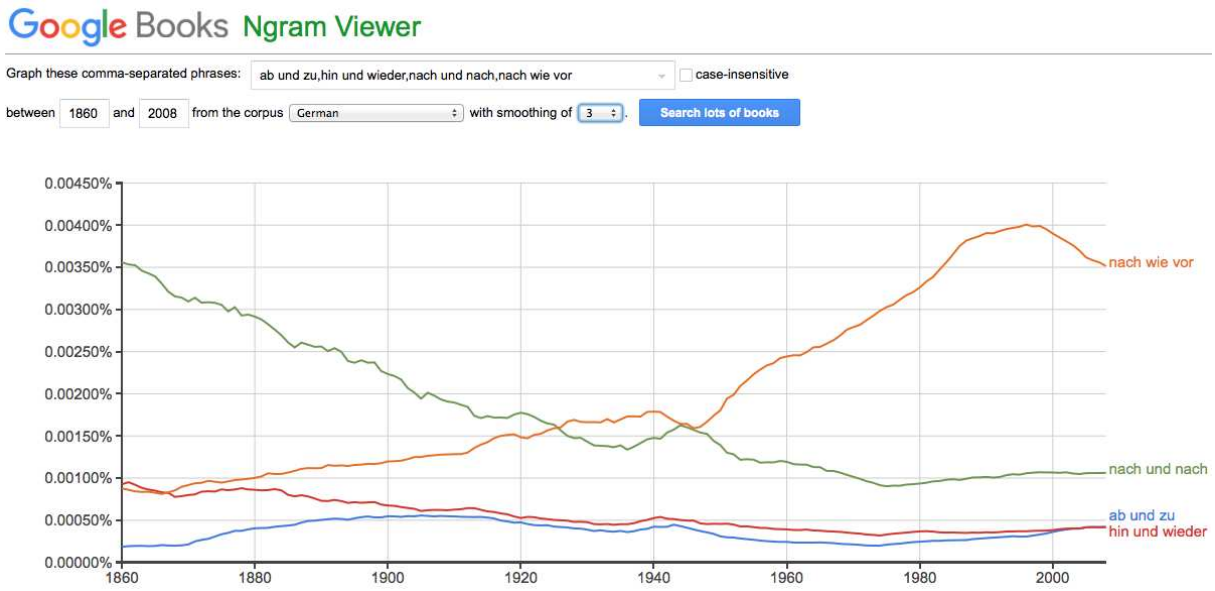

Figure 3: Google n-gram statistics showing the frequency development of four bi-particle adverbs over the last 150 years.

tics) that need special treatment in natural language processing. In order to detect the whole range of these adverbs we computed collocation scores for all patterns with words that are tagged as non-inflected adjectives (ADJD), adverbs (ADV, PAV, PWAV), prepositions (APPR), and separated prefixes (PTKVZ) in coordinated constructions with the conjunctions *als, oder, und, wie* (KON, KOKOM).

In this way we found fixed expressions like *fix und fertig; klipp und klar* (EN: wiped out; concisely) at the top of the list, but also pairs that stress opposition *hüben und drüben* (EN: here and there) or reinforce the meaning through synonym repetition *nie und nimmer* (EN: never ever) or reduplication *dunkler und dunkler* (EN: darker and darker)[2]. In addition we find pairs that form idiomatic adverbials within larger expressions *(über) kurz oder lang; (mehr/eher) schlecht als recht* (EN: sooner or later; badly). We also noted that bi-particle adverbs may also involve truncated words (tagged as TRUNC) as first conjunct as in *niet- und nagelfest, sang- und klanglos* (EN: nailed down; quietly).

In conclusion, bi-particle adverbs are an understudied category among multi-word expressions which deserves a lot more attention. These adverbs cover the whole spectrum of idiomaticity and can only be interpreted correctly when their collocation strength is appropriately considered.

4 Evaluation of Bi-particle Adverb Recognition on Separable Prefix Verbs

After automatic correction of the PoS tags in the bi-particle adverbs in table 1 we observe improved precision in the re-attachment of separated verb prefixes with 7600 prefix + verb combinations. We manually checked the re-attached prefix *nach* and found 79 cases with 1 error left. This error is due to a missed sentence boundary and a PoS error in a sentence-initial verb. Overall, we observe 47 removed prefix-verb combinations and 16 new prefix-verb combinations. All these changes are correct.

Recall of the re-attachment of separable verb prefixes is more difficult to determine. We see that there are still 1388 particles that are tagged as separated verb prefixes which we were unable to re-attach. We find 590 cases with a combination of prefix + verb which is not licensed through our list of separable prefix verbs, and 798 separated prefixes for which we do not find a full verb in the sentence. Most of these cases are PoS tagging

[2]English also features reduplications in adverbs: *again and again, more and more, neck and neck.*

errors either of the particle or the verb. For example, we have seen some PoS errors where the finite verb is mistakenly tagged as infinitive (the 1st and 3rd plural present tense forms of German verbs are homographic with the infinitive). The presence of a separated prefix indicates that the verb must be finite, and we could use that information to correct the verb's PoS tag if we trust the prefix tag more than the verb tag. This is currently not implemented.

For the unattachable words that are tagged as separated prefixes we found it to be advantageous to automatically correct their PoS tag to adverbs (ADV) for a list of 32 possible prefixes which often function as adverbs such as *empor, nahe, vorbei* (EN: upward, near, past). This correction step solves about half the cases where the PoS tagger assigned the tag "separated prefix" (PTKVZ) but we were unable to re-attach the word to a verb.

5 Related work

Lüdeling (2001) presents an in-depth study of the linguistic and corpus linguistic properties of German particle verbs. Stefan Müller (1999) discusses how to integrate German particle verbs into a comprehensive HPSG grammar whereas Forst et al. (2010) discuss the same for large LFGs. For both grammars it is unclear to what extent they could be used to annotate large corpora.

Hoppermann and Hinrichs (2014) introduce an approach to model particle verbs in their large German WordNet. Versley et al. (2010) have developed an approach for lemma disambiguation in German to serve the TüBa-D/Z treebank. In a recent publication Dewell (2015) investigates the semantics of selected German verb prefixes, both separable and inseparable ones.

Nießen and Ney (2000) report on early experiments to prepend German prefixes to the verbs for statistical machine translation into English. 14 years later Schottmüller (2014) still deals with separated verb prefixes in MT for the same language pair. She suggests to substitute German prefix verbs with synonymous inseparable verbs (e.g. substitute *fängt ... an* with *beginnt* (EN: to begin)) in order to improve translation quality. She demonstrates that current MT systems like Google Translate and Bing Translator still have problems with separated verb prefixes and produce better translations for sentences with synonymous non-separable verbs.

Related to our approach of the annotation of German prefix verbs is (Bott and Schulte im Walde, 2015) who present features to predict the compositionally of German particle verbs. Also similar is (Fritzinger, 2010) who uses parallel texts to detect German verb + prepositional phrase MWEs via automatic word alignment.

However, to the best of our knowledge, there is no literature on the interdependence between the recognition of multi-word adverbs and the analysis of separable prefix verbs. There is also no repository of German multi-word adverbs (unlike in French (Laporte and Voyatzi, 2008) and some other languages).

(Nagy and Vincze, 2014) present a method for the detection of verb-particle constructions in English (e.g. *to eat up, to take off*). They argue that a parser should be trained on a data set that includes specific annotation for verb-particle constructions.

Gereon Müller (1997) presents a detailed study of binomial constructions in German (e.g. *Fug und Recht, samt und sonders*) which includes bi-particle adverbs. He is particularly interested in order constraints (e.g. **Recht und Fug, *sonders und samt*) of the constructions. These constraints also hold for the bi-particle adverbs: **vor wie nach, *wieder und hin, *zu und ab* are not possible. Müller also offers a four level system of semantic opaqueness which would see the bi-particle adverb *hin und wieder* in class 1 (meaning is not compositional) and *auf und ab* in class 4 (meaning is compositional, but ordering constraints hold). He elaborates that end rhyme, alliteration (*ab und an*) and assonances (the repetition of vowel sounds to create internal rhyming) are typical properties of binominal constructions.

6 Conclusion

We have introduced an efficient algorithm for the computation of full lemmas for German verbs with separated prefixes. Checking the algorithm against the relevant verbs in the TüBA-D/Z treebank revealed an accuracy of 96.8%.

We have shown that the correct identification and PoS tagging of German bi-particle adverbs increases the accuracy of the re-attachment of separated prefixes to verb lemmas. Furthermore it

improves the interpretation and analysis of the sentences, both for the multi-word adverbs and the verbs. We also believe that the correct identification of multi-word adverbs and prefix verbs will improve cross-lingual word alignment and subsequently machine translation. This will be our next area of investigation.

Acknowledgments

This research was supported by the Swiss National Science Foundation under grant 105215_146781 for "SPARCLING: Large Scale PARallel Corpora for LINGuistic Investigation" (2013-2017) a joint project with Marianne Hundt and Elena Callegaro at the English Department of the University of Zurich. A first version of this work was presented at the PARSEME COST Action meeting in Struga, Macedonia in March 2016 with support by the European Union. We also thank the anonymous reviewers for helpful comments on an earlier version of this paper.

References

Stefan Bott and Sabine Schulte im Walde. 2015. Exploiting Fine-grained Syntactic Transfer Features to Predict the Compositionality of German Particle Verbs. In *Proceedings of the 11th Conference on Computational Semantics*, pages 34–39, London.

Robert B. Dewell. 2015. *The Semantics of German Verb Prefixes*, volume 49 of *Human Cognitive Processing*. John Benjamins.

Martin Forst, Tracy Holloway King, and Tibor Laczkó. 2010. Particle verbs in computational LFGs: Issues from English, German, and Hungarian. In *Proceedings of the LFG10 Conference*, pages 228–248. CSLI Publications.

Fabienne Fritzinger. 2010. Using parallel text for the extraction of German multiword expressions. *Lexis. E-Journal in English Lexicology*, pages 23–40, April.

Christina Hoppermann and Erhard Hinrichs. 2014. Modeling prefix and particle verbs in GermaNet. In Heili Orav, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the Seventh Global Wordnet Conference*, pages 49–54, Tartu, Estonia.

Eric Laporte and Stavroula Voyatzi. 2008. An electronic dictionary of French multiword adverbs. In *Proc. of LREC*, Marrakech, Morocco.

Anke Lüdeling. 2001. *On Particle Verbs and Similar Constructions in German*. CSLI, Stanford.

Gereon Müller. 1997. Beschränkungen für Binomialbildungen im Deutschen. *Zeitschrift für Sprachwissenschaft*, 16(1):25–51.

Stefan Müller. 1999. Syntactic properties of German particle verbs. In *Sixth International Conference on HPSG–Abstracts. 04–06 August 1999*, pages 83–88, Edinburgh.

István Nagy and Veronika Vincze. 2014. VPCTagger: Detecting verb-particle constructions with syntax-based methods. In *Proceedings of Workshop on Multiword Expressions. EACL*, Göteborg.

Sonja Nießen and Hermann Ney. 2000. Improving SMT quality with morpho-syntactic analysis. In *Proc. of COLING*, pages 1081–1085, Saarbrücken.

Nina Schottmüller. 2014. Issues in translating verb-particle constructions from German to English. In *Proceedings of Workshop on Multiword Expressions*, Gothenburg.

Yannick Versley, Kathrin Beck, Erhard Hinrichs, and Heike Telljohann. 2010. A syntax-first approach to high-quality morphological analysis and lemma disambiguation for the TüBa-D/Z treebank. In *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories*, pages 233–244, Tartu, Estonia.

Martin Volk, Chantal Amrhein, Noëmi Aepli, Mathias Müller, and Phillip Ströbel. 2016. Building a parallel corpus on the world's oldest banking magazine. In *Proceedings of KONVENS*, Bochum.